DOCUMENT RESUME

ED 478 206 TM 035 073

AUTHOR	Lin, Jie; Wu, Fenglan
TITLE	Differential Performance by Gender in Foreign Language Testing.
PUB DATE	2004-03-00
NOTE	18p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 22-24, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Adults; Foreign Countries; Item Bias; *Language Tests; *Listening Comprehension Tests; *Second Language Learning; *Sex Differences
IDENTIFIERS	China; DIMTEST (Computer Program); English Proficiency Test; *Item Bundles; SIBTEST (Computer Program)

ABSTRACT

Understanding and accounting for gender performance differences on high stakes examinations has become a particular concern for educational researchers to ensure test fairness for all examinees. In the context of second/foreign language proficiency testing, research suggests that males and females do not react differently at the item level. However, as R. Nandakumar (1993) suggested, items with small but systematic differential item functioning (DIF) may very often go statistically unnoticed, but when combined, they may be detected at the bundle level. Thus a study of differential bundle functioning (DBF) becomes necessary in order to understand more fully the influence of gender on test performance, especially when important, although perhaps subtle, secondary dimensions associated with different testlets have been found in the Test of English as a Foreign Language. In this study of the English Proficiency Test in China, the computer program SIBTEST was used for DIF/DBF analyses and DIMTEST for dimensionality testing. Subjects were 3,160 males and 1,299 females. The results indicate that although the English Proficiency Test did not demonstrate much gender DIF, the SIBTEST and DIMTEST analyses identified and confirmed the presence of the bundle of listening comprehension obviously favoring females and the bundles of grammar and vocabulary and close favoring males slightly. (Contains 3 tables, 2 figures, and 24 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 478 206

Differential Performance by Gender in Foreign Language Testing

Jie Lin and Fenglan Wu

The Centre for Research in Applied Measurement and Evaluation

The University of Alberta

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Lin

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Poster for the 2003 annual meeting of NCME in Chicago.

TM035073

BEST COPY AVAILABLE

Abstract

Understanding and accounting for gender performance differences on high stakes examinations has become a particular concern for educational researchers to ensure test fairness for all examinees. In the context of second/foreign language proficiency testing, research (Ryan and Bachman, 1992) suggests that males and females do not react differently at the item level. However, as Nandakumar (1993) suggested, items with small but systematic differential item functioning (DIF) may very often go statistically unnoticed, but when combined, they may be detected at the bundle level. Thus, a study of differential bundle functioning (DBF) becomes necessary in order to more fully understand the influence of gender on test performance, especially when important, although perhaps subtle, secondary dimensions associated with different testlets have been found in the TOEFL (Dunbar, 1982; Hale, Rock, & Jirele, 1989; Mckinley & Way, 1992). In the present study of the English Proficiency Test in China, the computer program SIBTEST was employed for DIF/DBF analyses and DIMTEST for dimensionality testing. The results indicate that although the English Proficiency Test did not demonstrate much gender DIF, the SIBTEST and DIMTEST analyses identified and confirmed the presence of the bundle of listening comprehension obviously favouring females, and the bundles of grammar and vocabulary, and cloze favouring males slightly.

Introduction

Understanding and accounting for possible gender differences has become a particular concern for educational researchers to ensure test fairness for all examinees. In the context of second/foreign language proficiency testing, however, gender differences have only been explored to a limited degree. Ryan and Bachman (1992) studied the differential performance on two well-known international tests, the TOEFL (the Test of English as a Foreign Language) and the FCE (the First Certificate of English). Little evidence was found that males and females reacted differently at the item level to either test. Similar results were also reported when the reading comprehension testlet of the TOEFL was studied (Wainer & Lukhele, 1997). However, as Wainer and Lukhele (1997) suggested, "it is not sufficient to merely examine each item for DIF [differential item funtioning], but the testlet itself must be examined in its totality"(p. 753). Very often, items with small but systematic DIF may go statistically unnoticed, but when combined, they may be detected at the bundle level (Nandakumar, 1993). Thus, a study of differential bundle functioning (DBF) becomes necessary in order to more fully understand the influence of gender on test performance, especially when important, although perhaps subtle, secondary dimensions associated with different testlets have been found in the TOEFL test (Dunbar, 1982; Hale, Rock, & Jirele, 1989; Mckinley & Way, 1992). The purpose of the present study was to explore whether a DBF analysis would reveal more evidence of differential functioning than a DIF analysis alone in the English Proficienct Test in China. In addition, the presence of secondary dimensions was also investigated as a common explanation for the differential bundle functioning (Shealy & Stout, 1993).

Literature Review

A number of studies conducted in various contexts have confirmed the presence of gender-related differences in verbal ability and language use (Maccoby & Jacklin, 1974; Thorne *et al.*, 1983; Tannen, 1990). The consensus seems to be that females are superior to males in general verbal ability (Maccoby & Jacklin, 1974; Denno, 1982; Cole, 1997), but there is disagreement about which types of verbal ability shows gender differences. This is especially true when it comes to different language skills.

Hyde and Linn (1988) conducted a comprehensive meta-analytical study investigating gender differences in verbal ability. Among the 56 vocabulary studies included, six reported a significant difference in favour of males, while eight reported significant differences in favour of females. Generally the meta-analysis demonstrated no significant gender difference in vocabulary, although there was significant heterogeneity in the effect size. In terms of reading comprehension, five out of the 21 studies reported a significant difference in favour of males, while ten found significant differences in favour of females. Generally, females were found to have slight advantages in reading, speaking, writing, and general verbal ability, but the differences were so small that Hyde and Linn argued that gender differences in verbal ability no longer existed. Statistics from ACT of 2001 also showed no significant sex differences in English or reading, although the means of females were slightly higher than those of males (Zwick, 2002). In contrast, a gender study recently conducted by the Educational Testing Service (ETS) yielded completely different results. This comprehensive study (Cole, 1997) involved 400 tests and millions of students. It was reported that a language advantage for females had remained unchanged compared with 30 years ago. As indicated in Figure 1, female superiority in verbal ability ranged from noticeable differences in writing and language use to very small differences in reading and vocabulary reasoning. At the same time, however, evidence also suggests that males are superior in listening vocabulary, that is, comprehension of heard vocabulary in both first and second language contexts (Brimer, 1969; Boyle, 1987). In general, despite the female advantage in general verbal ability, there seems to be no agreement as to whether and to what degree gender differences exist in different types of verbal ability.

In the context of second language proficiency testing, gender differences have been examined only to a limited degree. Generally, little differential performance by gender has been found. According to Ryan and Bachman (1992), the TOEFL did not demonstrate gender DIF. Of a total of 140 test items, no items were classified as 'C'(large DIF, as explained later in the paper). Of the six level B (moderate) DIF items, four favoured males and two favoured females. When means of subtests were compared, no significant gender differences were found in listening, structure and written expression, or vocabulary and reading. Wainer and Lukhele (1997) also reported that the reading comprehension testlets of TOEFL showed essentially no differential functioning by gender.

Method

Instrument and sample

The English Proficiency Test (EPT), one of the largest standardized English tests in China, was analyzed in the present study. The EPT is mainly used for assessing the English proficiency of adults who plan to seek further studies abroad at public expense. The subjects were typically university graduates with several years of work experience. Modelled after the TOEFL, the EPT includes Listening Comprehension (30 items), Grammar and Vocabulary (40 items), Cloze (20 items), Reading Comprehension (30 items), and Writing (1 item). In this study, all 120 multiple-choice items (the first four subtests) from the 1999 administration were examined. The sample included 3160 males and 1299 females.

Procedures

Differential item/bundle functioning (DIF/DBF)

Differential item functioning (DIF) analysis is a procedure often used to identify items that function differently between different groups, and thus help monitor the validity and fairness of tests. It is based on the assumption that test takers who have similar knowledge (based on total test scores) should perform in similar ways on individual test questions regardless of their sex, race, or ethnicity. Differential Item Functioning (DIF) occurs when an item is substantially harder for one group than for another group after the overall differences in knowledge of the subject tested are taken into account. Once the DIF items are detected statistically, there is a need for substantive interpretation to determine whether the items display bias or impact. If the item is biased, which unfairly favours one group of examinees over another, the item should be deleted or revised. If the item demonstrates impact, which reflects the actual difference in knowledge between the groups on the construct of interest, the item should be retained but further investigation may be necessary to explore why one group scored higher for this item.

Differential bundle functioning (DBF), a natural extension of DIF, examines the differential functioning of interpretable bundles of items instead of an individual item. The advantages of DBF lies in its increased power, more effectively controlled Type I error, and its ability to offer insight into DIF amplification (Bolt, 2002). Items with small but systematic DIF may go statistically unnoticed, but when combined, they may be detected at the bundle level

(Nandakumar, 1993). A bundle is a suspect subtest that is presumed to measure the primary dimension and a common secondary dimension, whereas the matching or valid subtest is believed to measure only the primary dimension. Once a bundle is flagged for DBF, there is also a need for substantive interpretation to determine whether the bundle displays bias or impact.

SIBTEST

The simultaneous item bias test (SIBTEST) implements a nonparametric statistical method of assessing DIF/DBF in an item or bundle of items based on Shealy-Stout's (1993) multidimensional model for DIF. The basic assumption is that multidimensionality produces DIF/DBF. SIBTEST detects bias by comparing the responses of examinees in the reference and focal groups that have been allocated to bins using their scores on a "matching subtest" (Stout & Roussos, 1995). The matching subtest is a subset of items that, ideally, are known to be unbiased. Roussos and Stout (1996) proposed the following guidelines for SIBTEST to classify DIF on a single item: (a) negligible or A-level DIF: Null hypothesis is rejected and the absolute value of beta-uni < 0.059; (b) moderate or B-level DIF: Null hypothesis is rejected and 0.059 =< the absolute value of beta-uni < 0.088; and (c) large or C-level DIF: Null hypothesis is rejected and the absolute value of beta-uni >= 0.088. For DBF, however, no guidelines exist for classifying the beta-uni values.

A four-step procedure (Gierl et al., 2001) was used to identify dimensions, if any, for which there were gender differences. First, the amount of DIF for each test item was obtained using SIBTEST (Stout & Roussos, 1995), and all items with B/C-level DIF were identified. Second, items were grouped by the four multiple-choice subtests of the EPT, and the beta-uni values for the items within each group were graphed. Third, interpretable bundles were identified by visually examining the graph and looking for groups of items that consistently favoured females or males. Fourth, the identified bundles were tested using the remaining items as the matching subtest after deleting items that displayed the most DIF, C-level DIF.

DIMTEST

To confirm the presence of secondary dimensions as identified in the SIBTEST analyses, DIMTEST analyses were conducted. A common explanation for the occurrence of DIF/DBF is the measurement of a nuisance dimension(s) unrelated to the primary dimension that is intended to be measured (Shealy & Stout, 1993; Roussos & Stout, 1996). While SIBTEST estimates the

amount of DIF/DBF beta-uni index, DIMTEST provides more direct evidence about a common source of DIF/DBF: multidimensionality. The DIMTEST statistic T and corresponding p-values are provided in the output. In this study, the DIMTEST analyses contained the same bundles as the studied and matching subtests in the SIBTEST analysis.

Results

Psychometric characteristics

The psychometric characteristics on the English Proficiency Test for males and females are summarized in Tables 1 and 2. Based on the total mean scores, there was no significant difference between the male and female examinees, although males did slightly better than females. This is an advantage for the present study in that the more similar the groups, the more accurate the DIF detection (Hambleton et al., 1993). The mean differences between males and females in each of the four sections were also tested using *t*-tests. The results indicated that females did significantly better than males in listening comprehension, while males outperformed females in both cloze, and grammar and vocabulary. When combined together, it is not surprising that there was no overall difference between male and female examinees on the English Proficiency Test.

SIBTEST results

The SIBTEST results did not show much gender DIF (see Table 3). Of the 120 items, two items (2%) exhibited C-level DIF and 11 items (9%) exhibited B-level DIF, while the majority (89%) exhibited no DIF. After controlling for ability, all four DIF items (including Levels B and C) in listening were found to be easier for females. The three grammar and vocabulary DIF items, and the two cloze DIF items were easier for males. In reading, however, three items favoured males while one favoured females.

Figure 2 provides the graphical presentation of the DIF effect size measure for each item by category. The x-axis represents the four categories and the y-axis represents the beta value for each item. Positive beta values favour males, while negative values favour females. This graph suggested that one bundle, listening, favoured females, while the cloze, and grammar and vocabulary bundles seemed to favour males slightly.

Further, the interpretable bundles were tested using SIBTEST. The matching or valid subtest included all the items from the remaining bundles except the C-level DIF item from reading comprehension. The DBF analysis (see Table 3) provided statistical evidence for the female advantage in listening ($\hat{\beta}_{uni}$ = -1.01), and a slight male advantage in Cloze ($\hat{\beta}_{uni}$ = 0.348) and Grammar and Vocabulary ($\hat{\beta}_{uni}$ = 0.667). There was no difference for the fourth category, Reading Comprehension.

DIMTEST results

The three bundles identified in the DBF analyses were further tested using DIMTEST. The DIMTEST statistics (see Table 3) showed that the listening subtest was by far the most distinct in dimensionality while the other two subtests, to a lesser degree, were also associated with secondary dimensions. Thus, the DIMTEST analyses confirmed the SIBTEST DBF results.

Discussion and Conclusions

This present study investigated gender differences in the English Proficiency Test in China, as one of the ways to ensure test fairness. In particular, results from DBF and DIF analyses were compared to see whether more evidence of differential functioning would be revealed in DBF analyses. In addition, this study examined whether secondary dimensions were present as a common explanation for the differential bundle functioning (Shealy & Stout, 1993).

The descriptive statistics indicated that there was no significant difference in overall English proficiency between the two groups, which seemed consistent with some previous findings: gender differences in verbal ability no longer exist (Hyde and Linn, 1988). Nevertheless, it should be pointed out that only receptive skills were compared in this study. What were excluded were writing and speaking skills in which females were often found to have an advantage in (Cole, 1997). In terms of their subtest performance, females had a higher mean in listening comprehension, which contradicted the findings of a male advantage in listening vocabulary (Brimer, 1969; Boyle, 1987). Males had a slight advantage in cloze, and grammar and vocabulary. It should be noted that although these differences were statistically significant, they were quite small in value, ranging from .35 to .73. The significant results may very likely be attributed to the large sample size. In general, it seems fair to say that the findings from this study

are to a certain degree consistent with Ryan and Bachman's (1992) assertions that no gender differences existed on TOEFL in any of the subtests.

Second, DIF analyses of individual items showed that the test did not demonstrate much gender DIF. Of the 120 items, 89% of items were DIF free, with eight DIF items favouring males and five favouring females. If the study had ended here, probably no gender differences would have been noticed in this test. However, the situation was different when bundle DIF came into play. DBF analyses demonstrated that the bundle of listening comprehension favoured females systematically, while the bundles of grammar and vocabulary, and cloze favoured males slightly.

Third, DIMTEST confirmed the dimensional distinctness of listening comprehension, grammar and vocabulary, and cloze from the rest of the test. This result is quite consistent with some previous findings on the TOEFL, after which the English Proficiency Test was modeled. Factor analytic research on the TOEFL seems to lead to the conclusion that the test measures primarily one factor, and at least, the TOEFL was unidimensional enough for the use of univariate item response theory (IRT) to be efficacious (Wainer and Lukhele, 1997). Nevertheless, with its three sections: listening comprehension, structure and written expression, vocabulary and reading comprehension, the TOEFL measures a variety of content areas and cognitive processes. It is thus reasonable to expect to find at least some empirical evidence of these dimensions in examinee response data. For instance, Dunbar (1982) found evidence of four factors: one general factor and one secondary factor associated with each of the three sections. Hale et al.,(1988) and Hale, Rock, and Jirele (1989) suggested a consistent two-factor structure of the TOEFL test, one related to listening comprehension, and one related to the remainder of the test. In a more recent study, Mckinley and Way (1992) applied both unidimensional IRT and multidimensional IRT to investigate the possible secondary TOEFL ability dimensions. They found that the TOEFL test was characterized by essentially three latent ability dimensions, a general ability dimension, a secondary ability dimension measuring listening comprehension, and a secondary ability dimension measuring a combination of structure and written expression and vocabulary and reading comprehension. In spite of the disparity among these studies, they all seem to agree on one point: there is a distinct dimension associated with listening comprehension. The present study provided further evidence for this dimensional distinctness of listening. The findings of secondary dimensions associated with grammar and vocabulary, and

cloze respectively are also somewhat consistent with Dunbar's sudy (1982).

Finally, some limitations of the current study must also be noted for future research. To begin with, the sample subjects were all from one province and not randomly selected. This lack of representativeness may affect the generalizability of the study to a certain degree. Next, it would be interesting to do a reliability check by reviewing the next administration of the test. If the same bundles are significantly flagged across both administrations, the findings of this study will be more conclusive. Further, more research on the dimensionality of the test using factor analysis, IRT and MIRT models will help to identify more interpretable dimensions of the test.

To sum up, in agreement with the suggestion of Wainer & Lukhele (1997), the DIF analyses of individual items showed that the EPT test did not demonstrate much gender DIF, but the DBF analyses did reveal differential performance: the bundle of listening comprehension favoured females, while Cloze, and Grammar and Vocabulary subtests tended to favour males, but to a smaller degree. Since Listening, Cloze, and Grammar and Vocabulary subtests each assesses different specific skills in addition to the general language ability, it is not surprising that DIMTEST confirmed the dimensional distinctiveness of these three subtests. Hence, in the context of foreign language testing, the present study demonstrated that a DBF analysis revealed more evidence about differential performance by gender than a DIF analysis. The presence of distinct dimensionality in turn helps explain the occurrence of the DBF. Next, to provide more insights into the nature of the content that may be related to differential functioning and to ensure test fairness, further substantive research and analysis would be needed.

References

Bolt, D. M. (2002). *Studying the DIF Potential of Nuisance Dimensions Using Bundle DIF and Multidimensional IRT Analyses.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Brimer, M. A. (1969). Sex differences in listening comprehension. *Journal of Research and Development in Education*, 9, 171-179.

Boyle, J. (1987). Sex differences in listening vocabulary. *Language Learning, 37*(2), 273-284.

Cole, N. S. (1997). *The ETS gender study: how females and males perform in educational setting.* Princeton, NJ: Educational Testing Service.

Denno, D. (1982). Sex differences in cognition: A review and critique of the longitudinal evidence. *Adolescence*, 17, 779-788.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*, 465-484.

Dunbar, S. B. (1982). *Construct validity and the internal structure of a foreign language test for several native language groups.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Gierl, M. J., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*, 26-36.

Gierl, M. J., Bisanz, J., Bisanz, G., & Boughton, K. (2002). *Identifying content and cognitive skills that produce gender differences in mathematics: a demonstration of the DIF analysis framework.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butler, F. A., & Oller, J. W. (1988). Multiple-choice cloze items and the Test of English as a Foreign Language (TOEFL Research Report No. 26). Princeton, NJ: Educational Testing Services.

Hale, G. A., Rock, D. A., & Jirele, T. (1989). Confirmatory factor analysis of the Test of English as a Foreign Language (TOEFL Research Report No. 32). Princeton, NJ: Educational Testing Services.

Hambleton, R. K., Clauser, B. E., Mazor, K. M., & Jones, R. W. (1993). Advances in detection of differentially functioning test items. *European Journal of Psychological Assessment, 9*, 1-18.

Hyde, J., & Linn, M. (1988). Gender differences in verbal ability: a meta-analysis. *Psychological Bulletin, 104*(1), 53-69

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.

McKinley, R. L., & Way, W. D. (1992). *The feasibility of modeling secondary TOEFL ability dimensions using multidimensional IRT models* (TOEFL Research Report N. 142). Princeton, NJ: Educational Testing Services.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 16*, 159-176.

Roussos, L, A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*, 215-230

Ryan, K., & Bachman, L. (1992). Differential item functioning on two tests of EFL proficiency. *Language testing, 9*(1), 12-29.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Stout, W., & Roussos, L. (1995). *SIBTEST manual*. University of Illinois: Department of Statistics, Statistical Laboratory for Educational and Psychological Measurement.

Tannen, D. (1990). *You just don't understand: women and men in conversation*. New York: William Morrow.

Thorne, B., Kramarae, C., & Henley, N. (Eds). (1983). *Language, gender and society*. Rowley, MA: Newbury House.

Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement, 57*(5), 741-759

Zwick, R. (2002). *Fair game? The use of standardized admissions tests in higher education*. New York: RoutledgeFalmer.

Table 1

Descriptive Statistics for the English Proficiency Test by Gender

		N	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
Females	Listening	1299	2.00	28.00	17.48	4.69	-.456	-.160
	Gram.& Voca.	1299	5.00	37.00	22.10	5.83	-.289	-.296
	Cloze	1299	1.00	17.00	8.97	2.73	.176	-.370
	Reading	1299	4.00	28.00	17.42	4.65	-.251	-.371
	Total	1299	21.00	103.00	65.97	14.71	-.336	-.231
Males	Listening	3160	2.00	29.00	16.75	4.63	-.290	-.433
	Gram.& Voca.	3160	5.00	39.00	22.65	5.89	-.272	-.374
	Cloze	3160	1.00	17.00	9.32	2.78	.055	-.268
	Reading	3160	3.00	30.00	17.57	4.72	-.302	-.381
	Total	3160	20.00	105.00	66.30	14.84	-.329	-.244

Table 2

t-test statistics for the English Proficiency Test by Gender

	Gender	Mean	Mean Difference	Std. Error	Sig.
Listening	female	17.4781	.731	.153	.000 *
	male	16.7475			
Gram. & Voca.	female	22.1008	-.553	.194	.004 *
	male	22.6538			
Cloze	female	8.9684	-.353	.091	.000 *
	male	9.3218			
Reading	female	17.4219	-.152	.155	.328
	male	17.5734			
Total	female	65.9692	-.327	.488	.502
	male	66.2965			

* $p < 0.01$.

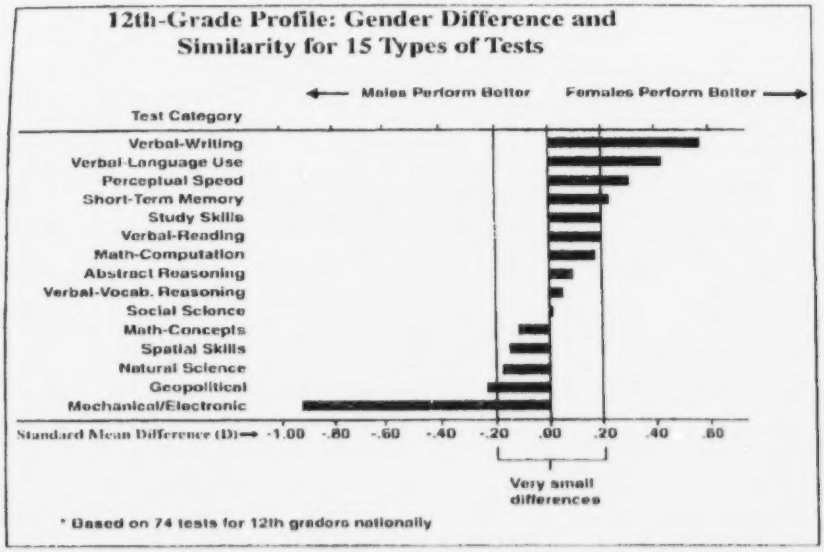
Table 3

Results of SIBTEST and DIMTEST

		SIBTEST						DIMTEST
		DIF				DBF		
	No. of items	No. of B-level DIF items	No. of C-level DIF items	Total No. of DIF items	Favouring	Beta Uni	Favouring	T-statitics
Listening	30	3	1	4	Females	-1.010*	Females	10.544*
Grammar & Vocabulary	40	3	0	3	Males	0.667*	Males	7.682*
Cloze	20	2	0	2	Males	0.348*	Males	5.404*
Reading	30	3	1	4	3 Males/ 1 Females	0.050	n/a	
Total	120	11	2	13	8 Males/ 5 Females			

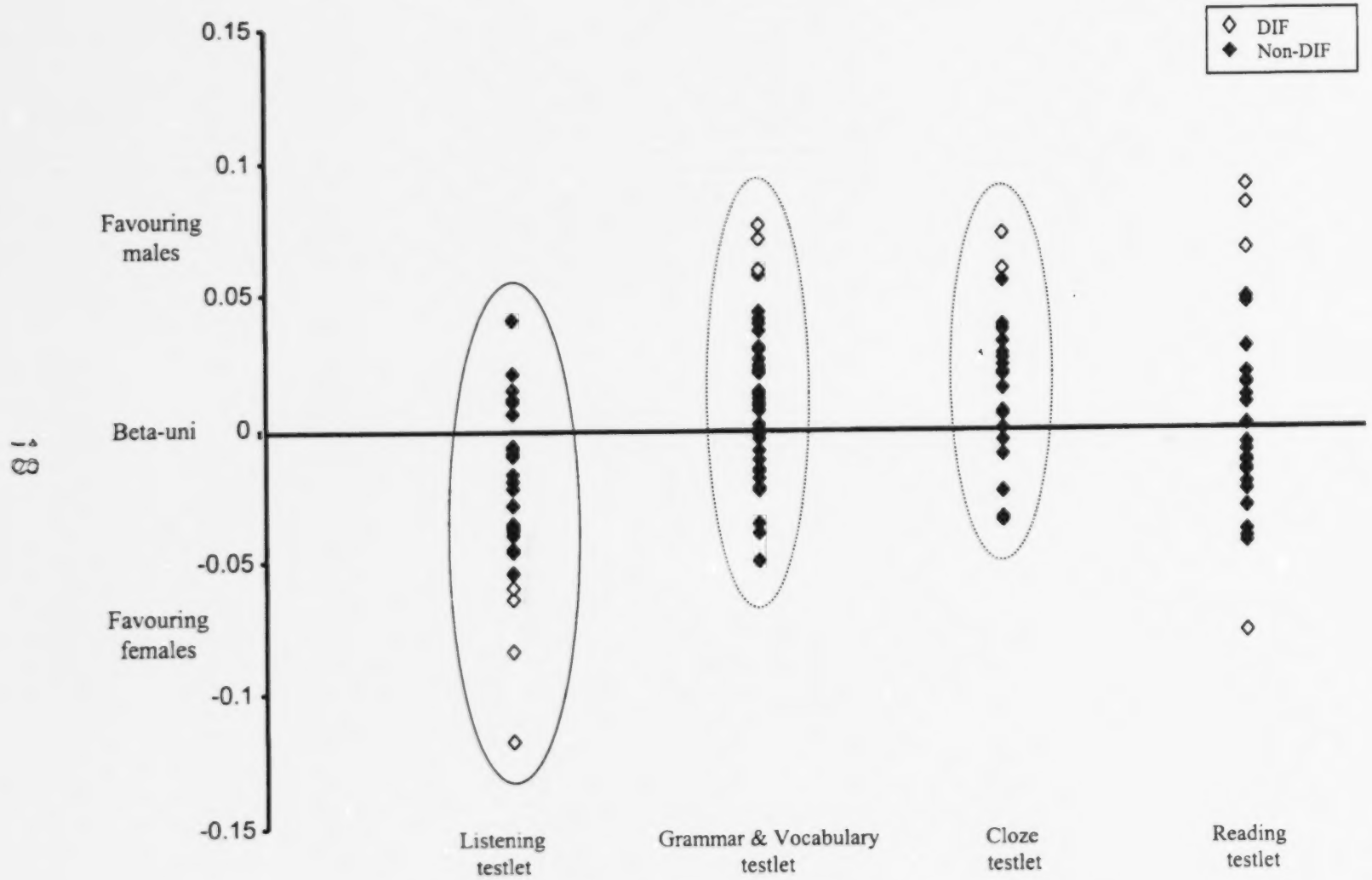
* $p < 0.01$

Figure 1



Source: Cole, 1997, p. 14.

Figure 2 DIF on the English Proficiency Test (Males vs. Females)



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM035073

I. DOCUMENT IDENTIFICATION:

Title:	Differential Performance by Gender in Foreign Language Testing
Author(s):	Jie Lin & Fenglan Wu
Corporate Source: University of Alberta	Publication Date: April, 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only.

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only.

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Jie Lin	Printed Name/Position/Title: Jie Lin / doctoral student
Organization/Address: University of Alberta Edmonton, Canada T6G 2G5 Centre for Research in Applied Measurement and Evaluation Dept. of Educational Psychology, 6-110 Education North	Telephone: (780) 492-5427 FAX: E-Mail Address: jie@ualberta.ca Date: June 18, 2003

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org

EFF-088 (Rev. 2/2001)